A Humor új Fo(r)mája

Novák Attila

MTA–PPKE Nyelvtechnológiai Kutatócsoport
Pázmány Péter Katolikus Egyetem Információtechnológiai és Bionikai Kar
1083 Budapest, Práter utca 50/a
novak.attila@itk.ppke.hu

Kivonat: A MorphoLogic Humor morfológiai elemzőjéhez az utóbbi évtizedekben számos nyelven készült morfológiai adatbázis. Ezek közül némelyik igen jó lefedettséget és pontosságot ad, mások olyan nyelvekre biztosítják az automatikus morfológiai elemzés lehetőségét, amelyekre más hasonló erőforrás nem létezik. A Humor elemzőszoftver zárt licence azonban nem tette lehetővé ezeknek a nyelvi erőforrásoknak a szabad terjesztését. Ugyanakkor a Humor elemző implementációja nem teszi lehetővé az ismeretlen szavak elemzését (morphological guessing), valamint azt sem, hogy az egyes szavakhoz gyakorisági információt rendeljünk, vagy a modellt másképp súlyozzuk. Ezeket a problémákat úgy oldottuk meg, hogy a Humor morfológiai erőforrásait olyan véges állapotú leírássá konvertáltuk, amely mindezeket a problémákat megoldja és rendelkezik nyílt forráskódú implementációval is.

1 Bevezetés

A MorhpoLogic Humor elemzője [7] számára számos nyelvhez készült jó minőségű morfológiai adatbázis. Ezek között számos agglutináló nyelv szerepel: a magyar [5] mellett az következő kis finnugor nyelvek: a komi, az udmurt, a mezei mari, az északi manysi és néhány hanti dialektus [6]. Ezek mellett különböző indoeurópai nyelvekhez, a lengyelhez, az angolhoz, a némethez, a franciához és a spanyolhoz is készült Humor leírás. Ezeknek a morfológiáknak a többsége egy az elemző által használt formátumnál magasabb szintű redundanciamentes jegyalapú leírás használatával készült, amelyből a Humor adatbázis automatikusan jön létre [5, 6].

Az eredeti Humor elemzőalgoritmus nem alkalmas arra, hogy a szó végződése alapján olyan szavak lehetséges elemzéseit előállítsa, amelyeknek a töve nem szerepel az adatbázisában. Nem is lenne egyszerű az algoritmust úgy módosítani, hogy képes legyen ennek a feladatnak a megoldására. Egy ilyen ismeretlenszó-elemző integrálása az elemzőbe ugyanakkor igen hasznos eszköz lenne, hiszen minden szöveg sok olyan szóalakot tartalmaz, amelynek a töve nem szerepel az elemző szótárában.

Emellett nem lehetséges a morfológiai modellek súlyozása vagy gyakorisági információval való ellátása sem, amelyre szükség lenne ahhoz, hogy a morfológia közvetlenül alkalmas legyen adatvezérelt szövegnormalizálási feladatok (pl. automatikus helyesírás-javítás vagy beszédfelismerés) támogatására. Szintén hasznos lenne a modellek súlyozhatósága az ismeretlenszó-elemző által generált javaslatok sorrendezé-

séhez. Ezek mellett a Humor hátrányaként a morfológiaielemző-szoftver zárt licence említhető, amely nem teszi lehetővé ezeknek a nyelvi erőforrásoknak a szélesebb körben való terjesztését.

Ebben a cikkben bemutatjuk, hogy hogyan oldottuk meg a fenti problémákat a Humor formátumú morfológiai leírások forrásának véges állapotú leírássá alakításával, amelyek kompilálására és használatára nyílt forráskódú eszközök is rendelkezésre állnak. A véges állapotú reprezentáció használható végződésalapú ismeretlenszó-elemzésre, természetes megoldást kínál gyakorisági információ hozzáadására a modellhez, és lehetővé teszi a súlyozott hibamodellekkel való kompozíciót.

2 A Humor morfológiai elemző

A program 'item-and-arrangement' típusú elemzést hajt végre: egy szóalak lehetséges elemzéseit morfsorozatokként adja meg. A szót felépítő minden morfnak kiírja a felszíni és mögöttes alakját, valamint a kategóriáját (amely strukturált információt is tartalmazhat, de lehet belső szerkezet nélküli címke is).

Az elemző mélységi keresést végez az adott szóalakon a lehetséges elemzések után. Olyan morfokat keres a szótárában, amelyeknek a felszíni alakja illeszkedik a megadott szó még elemezetlen részére. A lexikon nemcsak morfokat, hanem morfsorozatokat is tartalmazhat, amelyeket az elemző így egy lépésben ismer fel.

Elemzés közben a program kétféle ellenőrzést hajt végre. Egyrészt lokális kompatibilitás-ellenőrzést végez az egymás mellett álló morfok között, másrészt azt is ellenőrzi, hogy az elemzést alkotó morfémák a nyelv lehetséges szókonstrukciói egyikét testesítik-e meg. Az utóbbi ellenőrzést a szónyelvtant leíró kiterjesztett véges állapotú automata bejárásával ellenőrzi.

A Humor elemző lexikai adatbázisa a morfémák allomorfjainak leírásából, a szónyelvtant leíró véges állapotú automatából és a szomszédos morfémák lokális kompatibilitás-ellenőrzéséhez használt kétféle adatszerkezetből áll. Ezek egyikét folytatási osztályok és bináris kompatibilitási mátrixok alkotják, amelyek az egymással összekapcsolható folyatatási osztályokat adják meg. A másik adatszerkezetet bináris tulajdonságvektorok és megszorításvektorok alkotják. Minden morf leírása tartalmaz egy jobb és egy bal oldali folytatási osztálycímkét, egy jobb oldali bináris tulajdonságvektort és egy bal oldali bináris megszorításvektort. Az utóbbi tartalmazhat olyan pozíciókat, amelyek nem számítanak az illeszkedés szempontjából.

A lokális kompatibilitás-ellenőrzés a következő módon történik: egy adott morf (tipikusan egy toldalék) akkor illeszkedik az előző morfhoz (tipikusan tőhöz), ha a tő jobb oldali tulajdonságai kielégítik a toldalék bal oldali megszorításait mind a bináris tulajdonságok, mind a releváns folytatási mátrix szempontjából.

A szószerkezet globális ellenőrzéséhez használt szónyelvtan-automata bináris kiegészítő állapotváltozókat is tartalmazhat a fő állapotváltozója mellet, amelyek segítségével az automata méretének robbanása nélkül írhatók le a nem szomszédos morfémák közötti megszorítások.

Mindezek mellett a morfológiai adatbázis tartalmaz egy olyan leírást is, amely egy a jobb oldali tulajdonságvektorok halmazáról a morfológiai kategóriacímkék halmazára történő leképezést definiál. Ezeket a címkéket használjuk a szónyelvtan-

automata éleinek címkéiként. Az adott morf fellapozását és a lokális kompatibilitás-ellenőrzést minden esetben egy a szónyelvtan-automatában végzett lépés is követ. Az adott lépés akkor lehetséges, ha az automata adott állapotában (beleértve a kiterjesztett állapotváltozók aktuális értékét is) van olyan kimenő él, amely az adott morf jobb oldali tulajdonságvektora által meghatározott morfológiaicímke-halmaz valamelyik elemével van címkézve, és nem tartozik egyéb olyan megszorítás az adott élhez, amely a kiterjesztett állapotváltozók aktuális értékével nem kompatibilis.

Az adatbázis nehezen lenne karbantartható közvetlenül abban a formában, amelyben az elemző használja, mert ez az adatbázis-reprezentáció redundáns, alacsony szintű és nehezen olvasható formátumú adatszerkezeteket tartalmaz. Ezen problémák megoldására szolgál az a nyelviadatbázis-leíró keretrendszer, amelynek segítségével az adatbázis magas szintű és redundanciamentes formában írható le, amelyet a keretrendszer automatikusan alakít át az elemző által használt formára. A nyelviadatbázis-leíró keretrendszer létrehozás után keletkezett morfológiai leírások már ennek a magasabb szintű formalizmusnak a használatával készültek.

A magas szintű leírás leképezéséhez a rendszer egy kódolási leírást használ, amely minden egyes elemi tulajdonsághoz, amely a magas szintű leírásban szerepel, megadja, hogy az milyen alacsony szintű adatszerkezetre képződjön le és hogyan. Egyes tulajdonságok a bináris tulajdonságvektorokra képeződnek le, a többi pedig együtt határozza meg a folytatási mátrixokat, amelyeket dinamikusan generál a rendszer.

3 Véges állapotú morfológiák

A legszélesebb körben használt véges állapotú morfológiai eszközkészlet a Xerox *xfst-lookup* párosa [2]. Az *xfst* compilerrel különböző formalizmusok alkalmazásával lehet számítógépes morfológiákat leíró véges állapotú transzducereket létrehozni, amelyek morfológiai elemzőként vagy generátorként való működtetésére a *lookup* program szolgál. A morfológiai leírások a Xerox formalizmusában egyrészt a morfémákat leíró lexikális adatbázisból, másrészt a morfofonológiát leíró szabályrendszerből állnak. A lexikon definiálására szolgál a *lexc* formalizmus, amelynek segítségével leírhatók és allexikonokba szervezhetők a morfémák, és a szónyelvtan folytatási osztályok segítségével adható meg. Egy *lexc* allexikon általában olyan absztrakt morfémaleírásokból áll, amely a lemma és a morfoszintaktikai címkék mellett a morféma általában fonológiailag absztrakt ábrázolását tartalmazza. Az utóbbinak a szövegekben ténylegesen előforduló felszíni alakokra vetítését egy fonológiai-helyesírási szabályrendszer végzi.

A fonológiai szabályok szekvenciális és párhuzamos szabályrendszerként is megfogalmazhatók. Az *xfst* formalizmus és compiler segítségével szekvenciális újraírószabály-rendszerek adhatók meg és alakíthatók véges állapotú transzducerekké, illetve komponálhatók egymással és a lexikont leíró, a *lexc* formátumú leírásból kompilált transzducerrel. Így egyetlen lexikális transzducer hozható létre amely közvetlenül leképezi a felszíni szóalakokat a lemmákból és morfoszintaktikai címkékből álló lexikai reprezentációkra. Egy hasonló compiler, a *twolc* használható a párhuzamos kétszintű megszorítások segítségével megadott morfológiai leírások kompilálására.

A Xerox véges állapotú eszközkészlete a Humor elemző szónyelvtan-automatájában használt kiterjesztett állapotváltozókhoz hasonló formalizmus segítségével ugyancsak lehetővé teszi az állapottér faktorizációját. Az erre szolgáló konstrukciót a Xerox terminológiájában 'flag diacritics'-nek hívják.

Bár a flag diacritics használata általában csökkenti az elemző sebességét, ez a konstrukció mégis nagyon hasznos lehet, mert használatával megelőzhető, hogy a transzducer mérete exponenciálisan felrobbanjon a morfológiában szereplő nem lokális megszorítások következtében. Emellett arra is használható, hogy akár a morfémák közötti lokális megszorításokat is a puszta folytatási osztályoknál kifejezőbb és olvashatóbb formában írjuk le. Az xfst tartalmaz egy olyan műveletet, amelynek segítségével az ilyen lokális megszorításokat megfogalmazó flagek az automataméret számottevő növekedése nélkül eliminálhatók, növelve ezzel az elemző sebességét.

A Xerox eszközkészlete igen erős formalizmust ad a bonyolult morfológiai szerkezetek leírására. Ezért nem voltak komoly kétségeink azzal kapcsolatban, hogy a Humor formalizmus felhasználásával implementált morfológiai leírások átalakíthatóak lesznek a véges állapotú leírásokká.

Ugyanakkor, bár a Xerox eszközeit kutatási célra hozzáférhetővé tették 2003-ban Beesley és Karttunen könyvének [2] publikálásakor, ezek két szempontból nem különböznek lényegesen a Humor elemzőtől: zárt forráskódúak és nem használhatóak súlyozott modellek létrehozására. Ugyanakkor az ismeretlenszó-elemzés problémájának megoldására alkalmasak. Szerencsére néhány évvel ezelőtt létrejöttek az *xfst* és a *lookup* nyílt forráskódú alternatívái. Ezen nyílt forráskódú eszközök egyike, a Foma [3] alkalmas az *xfst-lexc* formátumú morfológiai leírások kompilálására és működtetésére. Ez tehát lehetővé teszi a zárt forráskód okozta problémák kiküszöbölését. Ezen kívül a szintén nyílt forráskódú HFST-eszközkészlet [4] segítségével a Foma formátumú transzducerek OpenFST [1] formátumúakká konvertálhatók, amely implementáció viszont lehetővé teszi a súlyozott véges állapotú modellek létrehozását.

4 A Humor–lexc konverzió

Mivel a Humor formalizmusban leírt morfológiai modellek a morfológia teljes leírását tartalmazzák a morfofonológiával együtt, ezen leírások átalakításához nincs szükség sem szekvenciális *(xfst)* sem párhuzamos *(twolc)* szabályrendszer használatára. Az átalakításhoz kizárólag a *lexc* formalizmust használjuk.

Minden morf lexikai alakját és kategóriacímkéjét a morf *lexc* reprezentációjának lexikai, a felszíni alakját pedig a felszíni oldalára képezzük le. Az utóbbi a valódi felszíni alak, nem egy olyan absztrakt fonológiai reprezentáció, amit általában a *lexc* lexikonforrásokban használni szoktak. A felszíni és a lexikai alakban egymásnak megfelelő szimbólumok helyes egymáshoz rendeléséről a lexikonkonverter implementációja gondoskodik. A címkéket egyetlen többkarakteres szimbólumként ábrázoljuk.

A Humor leírásban folytatási osztályok, mátrixok, illetve bináris tulajdonság- és megszorításvektorok formájában megadott lokális morfszomszédossági megszorításokat közvetlenül *lexc* folytatási osztályokként ábrázoljuk. A leképezés egyszerű implementálásához a Humor lexikonokat generáló programot kiegészítettük egy

olyan kapcsolóval, amelynek megadása esetén a program olyan mátrixokat generál, amelyek önmagukban teljesen leírják a szomszédossági megszorításokat, így a vektorok a *lexc* lexikonokra való leképezés folyamán figyelmen kívül hagyhatóak. Minden morf lexc reprezentációjának előállításakor az allexikont, amelybe az adott morf kerül, a morf bal mátrixának neve és a bal oldali folytatásiosztály-kódja határozza meg. A lexc folytatási osztályát ugyanakkor a jobb oldali mátrixnév, folytatásiosztály-kód és a szónyelvtan-kategóriacímke együttesen határozza meg.

A Humor szónyelvtan legkönnyebben a flag diacritics formalizmus segítségével képezhető le a véges állapotú formalizmusra. A Humor automata fő állapotváltozóját egy flagre képezzük le, amelyet St-nek neveztünk el. Ugyanakkor a kiterjesztett állapotváltozók mindegyike egy-egy újabb flagre képeződik le. Hogy pontosan milyen flag diacritics élek kapcsolódnak egy-egy morfhoz, azt az adott morf szónyelvtan-kategóriája határozza meg.

A jobb oldali mátrixnév és folytatási kód alapján a morfémalexikonok bal oldalához a Humor mátrixokban leírt kompatibilitási viszonyokat közvetlenül leíró allexikonokon keresztül csatoljuk vissza a Humor szónyelvtan-automatát leíró flag élek jobb oldalát. Az St szónyelvtan-állapotflag eliminálható a leírásból, de ez az állapottér jelentős megnövekedésével járhat.

5 Eredmények

Az alábbiakban röviden összehasonítjuk magyar morfológiánk egy 144000 morfot tartalmazó változatának eredeti Humorral kompilált változatának és az átalakított xfst-vel kompilált változatnak futásimemória-igényét és elemzési sebességét. A véges állapotú lexikonból két változat eredményeit mutatjuk be. Az első változatból nem eliminátuk az St flaget, a másodikból igen.

1. táblázat: Egy 144000 morfból álló magyar morfológiai leírás Humor és xfst által kompilált változatának összehasonlítása

	Humor lexikon	lexc – St flaggel	lexc – St flag eliminálva
futási memória	3,3 MB	20,6 MB	38,5 MB
elemzési sebesség	4700 szó/s	12500 szó/s	33333 szó/s

A véges állapotúvá alakítás jelentősen növeli az elemző memóriaigényét (>11-szeresére), ugyanakkor jelentős elemzésisebesség-növekedéssel is jár (>7-szeressel). Az St flag eliminálása majdnem kétszeresére növeli a lexikon méretét, ugyanakkor igen jelentős sebességnövekedéssel is jár. A további flagek eliminálása nagyjából szintén kétszeresére növeli a véges állapotú lexikon méretét minden egyes eliminált flaggel. Ez emellett rendkívül hosszú kompilálási időhöz is vezet. Ugyanakkor gyakorlatilag semmilyen további pozitív hatással nincs az elemzési sebességre.

Köszönetnyilvánítás

Ez a munka részben a TÁMOP-4.2.1./B-11/2-KMR-2011-0002 és a TÁMOP-4.2.2./B-10/1-2010-0014 pályázatok támogatásával készült.

Hivatkozások

1. Allauzen, C., Riley, M., Schalkwyk, J., Skut, W., Mohri, M.: Openfst: a general and efficient weighted finite-state transducer library. In Proceedings of the Ninth International Conference on Implementation and Application of Automata (CIAA 2007) 11–23
2. Beesley, K. R., Karttunen, L.: Finite State Morphology. CSLI Publications, Ventura Hall (2003)
3. Huldén M.: Foma: a finite-state compiler and library. In: Proceedings of EACL (2009) 29–32
4. Lindén, K., Axelson, E., Hardwick, S., Pirinen, T.A., Silfverberg, M.: HFST - Framework for Compiling and Applying Morphologies. In Proc. SFCM (2011) 67–85
5. Novák A.: Milyen a jó Humor? In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003), Szegedi Tudományegyetem (2003) 138–145
6. Novák, A.: Language Resources for Uralic Minority Languages. Proceedings of the SALTMIL Workshop at LREC 2008, Marrakech (2008) 27–32
7. Prószéky, G., Kis, B.: A Unification-based Approach to Morpho-syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages. In: Proceedings of the 37th Annual Meeting of the ACL, College Park, Maryland, USA (1999) 261–268